doi: 10.1046/j.1529-8817.2003.00065.x

Tests for Candidate-Gene Association Using Case-Parents Design

G. Zheng[1,*], Z. Chen[2] and Z. Li[3]

[1] *Office of Biostatistics Research, National Heart, Lung and Blood Institute, Bethesda, MD, USA*
[2] *Department of Statistics and Applied Probability, National University of Singapore, Republic of Singapore*
[3] *Department of Statistics, George Washington University, Washington, DC, USA Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA*

Summary

In the case-parents design for testing candidate-gene association, the conditional likelihood method based on genotype relative risks has been developed recently. A specific relation of the genotype relative risks is referred to as a genetic model. The efficient score tests have been used when the genetic model is correctly specified under the alternative hypothesis. In practice, however, it is usually not able to specify the genetic model correctly. In the latter situation, tests such as the likelihood ratio test (LRT) and the MAX3 (the maximum of the three score statistics for dominant, additive, and recessive models) have been used. In this paper, we consider the restricted likelihood ratio test (RLRT). For a specific genetic model, simulation results demonstrate that RLRT is asymptotically equivalent to the score test, and both are more powerful than the LRT. When the genetic model cannot be correctly specified, the simulation results show that RLRT is most robust and powerful in the situations we studied. MAX3 is the next most robust and powerful test. The TDT is the easiest statistic to compute, compared to MAX3 and RLRT. When the recessive model can be eliminated, it is also as robust and powerful as RLRT for other genetic models.

Keywords : Genetic model, MAX, Restricted likelihood ratio test, Robust, TDT, Trio

Introduction

Schaid & Sommer (1993) introduced the conditional likelihood approach in terms of genotype relative risks for testing candidate-gene association using case-parents designs. In a case-parents design, a family is ascertained when an affected child (case) of the family is ascertained. Then the genotypes of the case and both parents are obtained. Without assuming Hardy-Weinberg Equilibrium (HWE), Schaid & Sommer (1993) derived the conditional probabilities of case genotypes given their parental mating types, and used them to construct efficient score statistics for testing association between a genetic marker and a disease under various genetic models. They showed that the transmission/disequilibrium test (TDT) of Spielman *et al.* (1993) coincides with the score statistic that is optimal for the additive genetic model. Among the six parental mating types: (i) $MM \times MM$, (ii) $MM \times MN$, (iii) $MM \times NN$, (iv) $MN \times MN$, (v) $MN \times NN$, and (vi) $NN \times NN$, where M and N denote, respectively, the disease and normal alleles, only (ii), (iv) and (v) are informative. Therefore, in case-parents designs, only the data from these three types of families will be used.

Denote the penetrances at the candidate-gene locus as $f_0 = \text{pr}(\text{case}|NN)$, $f_1 = \text{pr}(\text{case}|MN)$ and $f_2 = \text{pr}(\text{case}|MM)$. The genotype relative risks were defined by Schaid & Sommer (1993) as $r_1 = f_1/f_0$ and $r_2 = f_2/f_0$, where f_0 is taken as the reference penetrance. The null hypothesis of no association is $f_0 = f_1 = f_2$, i.e., $r_1 = r_2 = 1$. In terms of genotype relative risks, a genetic model is recessive (REC) if $r_1 = 1$, additive (ADD) if $r_2 = 2r_1 - 1$, multiplicative (MUL) if

[*] Correspondence: Gang Zheng, Ph.D., Office of Biostatistics Research, National Heart, Lung and Blood Institute. 6701 Rockledge Drive, MSC 7938, Bethesda, MD 20892, U.S.A. Tel: (301)435 1287; Fax: (301)480 1862. E-mail: zhengg@nhlbi.nih.gov

Table 1 Conditional probability of case genotype, given mating type for various genetic models using (δ_0, δ_1)

			pr(case genotype|mating type)				
Mating Type	Case Genotype	Count	General (δ_0, δ_1)	REC $\delta_0 = \delta_1$	ADD $\delta_1 = \frac{1+\delta_0}{2}$	MUL $\delta_0 = \delta_1^2$	DOM $\delta_1 = 1$
(ii)	*MM*	n_{22}	$\frac{1}{1+\delta_1}$	$\frac{1}{1+\delta_1}$	$\frac{1}{1+\delta_1}$	$\frac{1}{1+\delta_1}$	$\frac{1}{2}$
	MN	n_{21}	$\frac{\delta_1}{1+\delta_1}$	$\frac{\delta_1}{1+\delta_1}$	$\frac{\delta_1}{1+\delta_1}$	$\frac{\delta_1}{1+\delta_1}$	$\frac{1}{2}$
(iv)	*MM*	n_{42}	$\frac{1}{1+2\delta_1+\delta_0}$	$\frac{1}{1+3\delta_1}$	$\frac{1}{4\delta_1}$	$\frac{1}{(1+\delta_1)^2}$	$\frac{1}{3+\delta_0}$
	MN	n_{41}	$\frac{2\delta_1}{1+2\delta_1+\delta_0}$	$\frac{2\delta_1}{1+3\delta_1}$	$\frac{1}{2}$	$\frac{2\delta_1}{(1+\delta_1)^2}$	$\frac{2}{3+\delta_0}$
	NN	n_{40}	$\frac{\delta_0}{1+2\delta_1+\delta_0}$	$\frac{\delta_1}{1+3\delta_1}$	$\frac{2\delta_1-1}{4\delta_1}$	$\frac{\delta_1^2}{(1+\delta_1)^2}$	$\frac{\delta_0}{3+\delta_0}$
(v)	*MN*	n_{51}	$\frac{\delta_1}{\delta_1+\delta_0}$	$\frac{1}{2}$	$\frac{\delta_1}{3\delta_1-1}$	$\frac{1}{1+\delta_1}$	$\frac{1}{1+\delta_0}$
	NN	n_{50}	$\frac{\delta_0}{\delta_1+\delta_0}$	$\frac{1}{2}$	$\frac{2\delta_1-1}{3\delta_1-1}$	$\frac{\delta_1}{1+\delta_1}$	$\frac{\delta_0}{1+\delta_0}$

$r_2 = r_1^2$, and dominant (DOM) if $r_1 = r_2$. Note that the definition of genetic models is meaningful only under the alternative hypothesis. For arbitrary (r_1, r_2), the model is referred to as a general genetic model. Schaid & Sommer (1993) derived the conditional probabilities of case genotypes given parental mating types in terms of (r_1, r_2). In this paper, we define the genotype relative risks differently as $\delta_0 = f_0/f_2$ and $\delta_1 = f_1/f_2$ by taking f_2 as the reference penetrance for the reason that follows. Since M is the disease allele, $0 < f_0 \leq f_1 \leq f_2$ and hence δ_0 and δ_1 are bounded, that is, $0 \leq \delta_0 \leq \delta_1 \leq 1$. However, r_1 and r_2 do not share this boundedness property. Though the two definitions of genotype relative risks are equivalent in the sense that there is a one-to-one correspondence between them and that the statistics for testing the null hypothesis are the same by using either (r_1, r_2) or (δ_0, δ_1), the lack of boundedness might entail difficulties in the computation of the test statistics. In terms of (δ_0, δ_1), the null hypothesis becomes $H_0 : \delta_0 = \delta_1 = 1$. In the (δ_0, δ_1) plane, the genotype relative risks are constrained to the triangle $T = \{(\delta_0, \delta_1) : 0 \leq \delta_0 \leq \delta_1 \leq 1\}$. Any point in T except the point $(\delta_0, \delta_1) = (1, 1)$, which specifies the null hypothesis, corresponds to an alternative hypothesis.

Let n_2, n_4 and n_5 denote the numbers of ascertained families with parental mating type (ii), (iv) and (v), respectively. Let n_{ij} denote the number of cases that have a genotype with j disease alleles and are from family type i, $i = 2, 4, 5$, $j = 0, 1, 2$. Then $n_2 = n_{21} + n_{22}$, $n_4 = n_{40} + n_{41} + n_{42}$ and $n_5 = n_{50} + n_{51}$. Conditional on the parental mating types (ii), (iv) and (v) and n_2, n_4, n_5, (n_{22}, n_{21}) and (n_{51}, n_{50}) have binomial distributions, and (n_{42}, n_{41}, n_{40}) has a trinomial distribution. The numbers of cases, i.e., (n_{22}, n_{21}), (n_{42}, n_{41}, n_{40}) and (n_{51}, n_{50}) are conditionally independent given n_2, n_4, n_5. In Table 1, the conditional probabilities given in Schaid (1999) are re-expressed in terms of (δ_0, δ_1). The last four columns of Table 1, corresponding to the conditional probabilities of four genetic models, are obtained from the general genetic model.

Denote the probabilities of mating type (ii), (iv) and (v) by p_2, p_4 and p_5, respectively, which were given in Schaid (1999). In terms of (δ_0, δ_1), $p_2 = 2p^3q(1 + \delta_1)/R$, $p_4 = p^2q^2(1 + 2\delta_1 + \delta_0)/R$ and $p_5 = 2pq^3(\delta_0 + \delta_1)/R$, where p is the frequency of allele M, $q = 1 - p$ and $R = p^2 + 2\delta_1 pq + \delta_0 q^2$. In this paper, we assume that a large number of families with case are screened and $n = n_2 + n_4 + n_5$ informative families with case are obtained. The number n is referred to as the sample size of the case-parents design.

Various test statistics based on the conditional likelihood in terms of (r_1, r_2) have been studied in the literature. We review briefly these statistics and study the restricted likelihood ratio tests. Comparison of the restricted likelihood ratio tests with other tests in terms of their robustness and power is presented.

Review of Existing Test Statistics

We review briefly the tests available in the literature for detecting disease-gene association using case-parents designs. The tests can be divided into two classes.

One class of tests is for the situation that the genetic model is specified under the alternative hypothesis, the other for the situation that the genetic model cannot be correctly specified under the alternative hypothesis.

The Situation that the Genetic Model is Specified

Schaid & Sommer (1993) proposed efficient score test for each of the four genetic models mentioned before, that is, the recessive (REC), dominant (DOM), additive (ADD) and multiplicative (MUL) models. They derived the following score statistics, respectively, for recessive, dominant and additive models:

$$\begin{aligned} Z_{\mathrm{REC}} &= (4n_{22} + 4n_{42} - 2n_2 - n_4)/(4n_2 + 3n_4)^{1/2} \\ Z_{\mathrm{DOM}} &= (n_4 + 2n_5 - 4n_{40} - 4n_{50})/(3n_4 + 4n_5)^{1/2}, \\ Z_{\mathrm{ADD}} &= (n_{22} + 2n_{42} + n_{51} - n_{21} - 2n_{40} - n_{50})/ \\ &\quad \times (n_2 + 2n_4 + n_5)^{1/2}. \end{aligned}$$

They also pointed out that the score statistic for the multiplicative model is the same as that for additive model.

Note that the above three score statistics can be obtained as special cases from a general score statistic as follows. Under the alternative hypothesis $(\delta_0, \delta_1) \in T - \{(1, 1)\}$, let $\lambda = 1 - \delta_0$ and $x = (\delta_1 - \delta_0)/(1 - \delta_0)$, i.e., $\delta_0 = 1 - \lambda$ and $\delta_1 = 1 - \lambda(1 - x)$, where $0 \le x, \lambda \le 1$. This reparameterization establishes a one-to-one relationship between genotype relative risks (δ_0, δ_1) and (x, λ) under the alternative. With this reparameterization, the null and alternative hypotheses becomes $H_0 : \lambda = 0$ and $H_a : \lambda > 0$, respectively. By using the probabilities for the general model given in the fourth column of Table 1 with $\delta_0 = 1 - \lambda$ and $\delta_1 = 1 + \lambda(x - 1)$, the likelihood function is proportional to

$$L_1(\lambda, x) = \frac{(1 - \lambda)^{n_{40}+n_{50}}[1 + \lambda(x - 1)]^{n_{21}+n_{41}+n_{51}}}{[2 + \lambda(x - 1)]^{n_2}[4 + \lambda(2x - 3)]^{n_4}[2 + \lambda(x - 2)]^{n_5}}, \tag{1}$$

where the n_i's and n_{ij}'s are defined in the introduction section. For a fixed x, the score function and the Fisher information about λ evaluated at $\lambda = 0$ are given, respectively, by

$$\begin{aligned} \left.\frac{\partial \log L_1(\lambda, x)}{\partial \lambda}\right|_{\lambda=0} &= -(n_{40} + n_{50} - n_4/4 - n_5/2) \\ &\quad +(x - 1)(n_{21} + n_{41} + n_{51} - n/2), \\ -\mathrm{E}\left[\frac{\partial^2 \log L_1(\lambda, x)}{\partial \lambda^2}\right]_{\lambda=0} &= n_2(x - 1)^2/4 \\ &\quad +n_4\{1/4 + (x - 1)^2/2 - (2x - 3)^2/16\} + n_5\{1/ \\ &\quad \times 2 + (x - 1)^2/2 - (x - 2)^2/4\}, \end{aligned}$$

where $n = n_2 + n_4 + n_5$. The efficient score statistic is then given by

$$\begin{aligned} Z_x &= \frac{\partial/\partial \log L_1(\lambda, x)|_{\lambda=0}}{\left[-\mathrm{E}\{\partial^2/\partial\lambda^2 \log L_1(\lambda, x)\}|_{\lambda=0}\right]^{1/2}} \\ &= \frac{a + xb}{(dx^2 - 2ex + c)^{1/2}}, \end{aligned} \tag{2}$$

where $a = 3n_4 - 4(n_{40} + n_{41}) + 2n_2 - 4n_{21}$, $b = 4n_{21} + 4n_{41} + 4n_{51} - 2n$, $c = 4n_2 + 3n_4$, $d = 4n$, and $e = 4n_2 + 2n_4$. It is easy to verify that $Z_{\mathrm{REC}} = Z_0$, $Z_{\mathrm{DOM}} = Z_1$ and $Z_{\mathrm{ADD}} = Z_{1/2}$. When n_2, n_4 and n_5 are all large, the score test statistic Z_x has asymptotically a standard normal distribution under the null hypothesis. Thus, if $Z_x > z_{1-\alpha}$, the null hypothesis is rejected in favor of alternative hypothesis with the genetic model specified by x, where $z_{1-\alpha}$ is the upper α quantile of the standard normal distribution, and α is the probability of the Type I error.

Schaid (1999) also derived the likelihood ratio test (T_{LRT1}) for each of the four genetic models. Here, the subscript 1 in T_{LRT1} indicates the number of parameters with respect to which the likelihood is maximized.

The Situation that the Genetic Model is Unspecified

In this situation, two robust tests were studied for testing the null hypothesis of no association. The first test is the general likelihood ratio test (LRT) studied by Schaid (1999) which does not impose any restriction on the relative risks (r_1, r_2). We now give the general LRT statistic in terms of (δ_0, δ_1). From Table 1, the conditional likelihood function for a general genetic model is proportional to

$$L_2(\delta_0, \delta_1) = \frac{\delta_1^{n_{21}+n_{41}+n_{51}}\delta_0^{n_{40}+n_{50}}}{(1+\delta_1)^{n_2}(1+2\delta_1+\delta_0)^{n_4}(\delta_0+\delta_1)^{n_5}}. \tag{3}$$

Then LRT statistic is given by

$$T_{\mathrm{LRT2}} = 2\log\{L_2(\hat{\delta}_0, \hat{\delta}_1)/L_2(1,1)\}, \tag{4}$$

where $(\hat{\delta}_0, \hat{\delta}_1)$ is the maximum likelihood estimator (MLE) of (δ_0, δ_1). The MLE can be obtained by solving the likelihood equations $\partial/\partial\delta_0 \log L_2(\delta_0, \delta_1) = 0$ and $\partial/\partial\delta_1 \log L_2(\delta_0, \delta_1) = 0$. The explicit expressions for $\partial/\partial\delta_i \log L_2(\delta_0, \delta_1)$, $i = 0, 1$ are given in Appendix A. By the classical asymptotic theory, the LRT statistic T_{LRT2} follows asymptotically the Chi-square distribution with 2 degrees freedom under the null hypothesis.

The second test was studied by Zheng *et al.* (2002). When the genetic model is unspecified, Z_x given by (2) cannot be used, as x can not be estimated under H_0. In this situation, Zheng *et al.* (2002) studied a robust test based on the statistic MAX3 = $\max(Z_0, Z_{1/2}, Z_1)$. The critical value z^* such that $\mathrm{pr}_{H_0}(\mathrm{MAX3} < z^*) = 1 - \alpha$ has no closed form. It has to be obtained from simulation under the null hypothesis. Note that, for a large sample sizes, (Z_0, Z_1) has a joint bivariate normal distribution with mean zero, variances 1 and covariance $\rho_{01} = \mathrm{corr}_{H_0}(Z_0, Z_1)$. The expression of ρ_{01} was given in Zheng *et al.* (2002). Moreover, $Z_{1/2}$ can be expressed as a linear combination of Z_0 and Z_1. Thus, the asymptotic distribution of MAX3 can be simulated by generating independent bivariate normal vectors with zero mean vector and the given variance-covariance matrix.

Both T_{LRT2} and MAX3 seem to be robust (Schaid, 1999; Zheng *et al.* 2002). However, their powers under alternative hypotheses have not been compared yet.

Restricted Likelihood Ratio Tests

The general likelihood ratio test considered by Schaid (1999) ignores the fact that $0 \le \delta_0 \le \delta_1 \le 1$, which is intrinsic in the underlying genetic model. Intuitively, by ignoring this fact, the efficiency of the tests will be adversely affected. Therefore, it is more appropriate to consider a restricted version of the likelihood ratio test by taking into account this fact. It should be mentioned that the restricted likelihood ratio tests (RLRT) were applied to linkage analysis by several authors. Holmans (1993) considered the RLRT for the linkage analysis using affected sib-pairs. Kruse *et al.* (1997) and Knapp (1998) considered the RLRT for the linkage analysis using extreme discordant sib-pairs. Here, we explore the application of the RLRT for testing disease-gene association in case-parents designs.

When the Genetic Model is Specified

A genetic model here means that a given functional relationship between δ_0 and δ_1, $\delta_1 = g(\delta_0)$, e.g., $\delta_1 = (1+\delta_0)/2$ for the additive model. When the genetic model is specified, i.e., the g function is known, the RLRT has a general form

$$T_{\mathrm{RLRT1}} = 2\log\left\{\max_{0\le\delta_0\le 1} L_2(\delta_0, g(\delta_0))/L_2(1,1)\right\},$$

where the likelihood L_2 is given by (3). Applying the RLRT to the recessive ($\delta_0 = \delta_1$), additive ($\delta_1 = (1+\delta_0)/2$), multiplicative ($\delta_1 = \delta_0^{1/2}$), or dominant ($\delta_1 = 1$) models, the RLRT statistics are, respectively, given by

$$T_{\mathrm{RLRT1}}^{\mathrm{REC}} = 2\log\left\{\max_{0\le\delta_0\le 1} L_2(\delta_0, \delta_0)/L_2(1,1)\right\},$$

$$T_{\mathrm{RLRT1}}^{\mathrm{ADD}} = 2\log\left\{\max_{0\le\delta_0\le 1} L_2(\delta_0, (1+\delta_0)/2)/L_2(1,1)\right\}$$
$$= 2\log\left\{\max_{1/2\le\delta_1\le 1} L_2(2\delta_1 - 1, \delta_1)/L_2(1,1)\right\},$$

$$T_{\mathrm{RLRT1}}^{\mathrm{MUL}} = 2\log\left\{\max_{0\le\delta_0\le 1} L_2\left(\delta_0, \delta_0^{1/2}\right)/L_2(1,1)\right\}$$
$$= 2\log\left\{\max_{0\le\delta_1\le 1} L_2\left(\delta_1^2, \delta_1\right)/L_2(1,1)\right\},$$

$$T_{\mathrm{RLRT1}}^{\mathrm{DOM}} = 2\log\left\{\max_{0\le\delta_0\le 1} L_2(\delta_0, 1)/L_2(1,1)\right\}.$$

To find the above restricted MLE, one can solve $\partial L_2(\delta_0, g(\delta_0))/\partial\delta_0 = 0$ for unrestricted MLE (see Appendix B). If the unrestricted MLE is in the range, it is also restricted MLE; otherwise use the boundary value as restricted MLE. The distribution of T_{RLRT1} for each of the four genetic models under the null hypothesis is no longer asymptotically the chi-square distribution with 1 degree of freedom, χ_1^2, since the classical asymptotic theory on the likelihood ratio test does not apply here. From Self & Liang (1987), T_{RLRT1} follows asymptotically a mixture distribution, $(1/2)\chi_1^2 + (1/2)\chi_0^2$, where χ_0^2 is a degenerate distribution with all mass at 0.

When the Genetic Model is Unspecified

In the situation that the genetic model is unspecified, one has to maximize the likelihood function over the region T. The RLRT statistic for the general genetic model, denoted by T_{RLRT2}, is given by

$$T_{\mathrm{RLRT2}} = 2\log\left\{\sup_{(\delta_0,\delta_1)\in T} L_2(\delta_0,\delta_1)/L_2(1,1)\right\} \\ = 2\log\frac{L_2(\hat{\delta}_0^*,\hat{\delta}_1^*)}{L_2(1,1)}. \quad (5)$$

where $(\hat{\delta}_0^*,\hat{\delta}_1^*)$ is the restricted MLE of (δ_0,δ_1) in T.

To obtain $(\hat{\delta}_0^*,\hat{\delta}_1^*)$, the unrestricted MLE $(\hat{\delta}_0,\hat{\delta}_1)$ is obtained first. When $(\hat{\delta}_0,\hat{\delta}_1)$ is contained in T, $(\hat{\delta}_0^*,\hat{\delta}_1^*) = (\hat{\delta}_0,\hat{\delta}_1)$ is the restricted MLE. When $(\hat{\delta}_0,\hat{\delta}_1)$ is not contained in T and the Hessian matrix is positive definite, $(\hat{\delta}_0^*,\hat{\delta}_1^*)$ must be on the boundaries of T, which are specified by $\delta_0 = 0, \delta_0 = \delta_1$ and $\delta_1 = 1$, respectively. Note that $L_2(\delta_0,\delta_1)$ is zero when $\delta_0 = 0$. Thus $(\hat{\delta}_0^*,\hat{\delta}_1^*)$ must be on the other two boundaries, corresponding to the recessive and dominant models, respectively. Thus, applying RLRT, we can find restricted MLE from recessive and dominant models. Let $\hat{\delta}_{\mathrm{REC}} = \mathrm{argmax}_{0\le\delta\le 1} L_2(\delta,\delta)$, and $\hat{\delta}_{\mathrm{DOM}} = \mathrm{argmax}_{0\le\delta\le 1} L_2(\delta,1)$. Then $(\hat{\delta}_0^*,\hat{\delta}_1^*) = (\hat{\delta}_{\mathrm{REC}},\hat{\delta}_{\mathrm{REC}})$ if $L_2(\hat{\delta}_{\mathrm{REC}},\hat{\delta}_{\mathrm{REC}}) \ge L_2(\hat{\delta}_{\mathrm{DOM}},1)$, and $(\hat{\delta}_0^*,\hat{\delta}_1^*) = (\hat{\delta}_{\mathrm{DOM}},1)$ otherwise. However, it is not easy to show analytically that the Hessian matrix is positive definite for any sample size n and any (δ_0,δ_1) contained in T, although, given the data, the Hessian matrix can be evaluated for any (δ_0,δ_1) contained in T numerically. An alternative approach to find the restricted maximum likelihood estimators without evaluating the Hessian matrix is the grid search of $(\delta_0,\delta_1) \in T$ with a step size, say, 0.005 for both δ_0 and δ_1, maximizing the likelihood function $L_2(\delta_0,\delta_1)$ over T. This is the advantage of using bounded genotype relative risks (δ_0,δ_1).

The null distributions of (4) and (5) are different. Under the null hypothesis, $(\delta_0,\delta_1) = (1,1)$ is on the boundary of the triangle T while it is the inner point of the space of genotype relative risks without constraints. Thus, T_{RLRT2} does not follow a chi-square distribution with two degrees of freedom. From Self & Liang (1987), T_{RLRT2} follows a mixture distribution, $(\phi/(2\pi))\chi_2^2 + (1/2)\chi_1^2 + (1/2 - \phi/(2\pi))\chi_0^2$, where ϕ is given by (see Appendix A)

$$\cos\phi = \mathrm{corr}_{H_0}(Z_0, Z_1) = \frac{q_4}{(4q_5+3q_4)^{1/2}(4q_2+3q_4)^{1/2}}, \quad (6)$$

where q_2, q_4 and q_5 are the observed proportions of total n samples from mating type (ii), (iv) and (v), respectively, and $q_i \to p_i$ for $i = 2, 4, 5$, as $n \to \infty$. When the significance level $\alpha = 0.05$, the critical value for T_{LRT2} given by (4) is 5.991, i.e., the null hypothesis is rejected when $T_{\mathrm{LRT2}} > 5.991$. The critical value for T_{RLRT2} can be simulated under the null hypothesis using (6). For example, given $n = 150$ informative families of mating type (ii), (iv) and (v), the critical values for T_{RLRT2} is 4.151, 4.138 and 4.172 when the frequency of allele M is 0.05, 0.20 and 0.50, respectively. The critical values may also be obtained using Splus functions.

Simulation Results

The empirical power of different tests is compared under various genetic models by simulations. The simulations are conditional on the sample size $n = n_2 + n_4 + n_5 = 150(200)$ families of mating type (ii), (iv) and (v) with case. Given the M-allele frequency p and genotype relative risks δ_0 and δ_1, the conditional expectation of the sample size for each informative mating type is $E(n_i|n) = np_i/(p_2+p_4+p_5)$ for $i = 2, 4, 5$, where p_i are defined in Introduction under HWE. When HWE does not hold, we assume the data are drawn from a mixture of two populations with different allele frequencies p^* and p^{**}. Within each sub-population, assume HWE holds. Thus, the conditional expectation of the sample size for each mating type within each sub-population can be calculated as before using p^* or p^{**}. Also assume that genotype relative risks in each sub-population are the same as those in the general population.

In the simulation, given p (or p^{**} and p^{**}), n_i, and genotype relative risks, case genotypes were generated under the alternative hypothesis using the conditional probabilities given in Table 1. The counts n_{ij}, $i = 2, 4, 5$

Table 2 Empirical power when the genetic model is known and HWE holds

True model	p	Z_x	T_{LRT1}	T_{RLRT1}
$\alpha = 0.05$ and $n = 150$				
REC ($\delta_0 = 0.4, \delta_1 = 0.4$)	0.05	0.3251	0.3250	0.3250
	0.20	0.9011	0.8390	0.9011
	0.50	0.9989	0.9968	0.9989
ADD ($\delta_0 = 0.6, \delta_1 = 0.8$)	0.05	0.5787	0.4432	0.5541
	0.20	0.5228	0.4707	0.5764
	0.50	0.5267	0.5038	0.5509
MUL ($\delta_0 = 0.49, \delta_1 = 0.7$)	0.05	0.7685	0.6599	0.7685
	0.20	0.7593	0.6548	0.7593
	0.50	0.7940	0.7024	0.7940
DOM ($\delta_0 = 0.7, \delta_1 = 1.0$)	0.05	0.6566	0.5923	0.6566
	0.20	0.6778	0.5399	0.6778
	0.50	0.4273	0.3356	0.5183
$\alpha = 0.01$ and $n = 200$				
REC ($\delta_0 = 0.4, \delta_1 = 0.4$)	0.05	0.2267	0.2267	0.2267
	0.20	0.8887	0.8322	0.8887
	0.50	0.9991	0.9950	0.9991
ADD ($\delta_0 = 0.6, \delta_1 = 0.8$)	0.05	0.4253	0.3215	0.4122
	0.20	0.3697	0.3469	0.4149
	0.50	0.3699	0.3528	0.3929
MUL ($\delta_0 = 0.49, \delta_1 = 0.7$)	0.05	0.5984	0.4899	0.5984
	0.20	0.6344	0.5844	0.6344
	0.50	0.7008	0.6577	0.7401
DOM ($\delta_0 = 0.7, \delta_1 = 1.0$)	0.05	0.5360	0.4461	0.5360
	0.20	0.5323	0.4147	0.5323
	0.50	0.2237	0.1680	0.2237

and $j = 0, 1, 2,$ were obtained. Table 2 reports the empirical power of score statistics, T_{LRT1} and T_{RLRT1} under HWE when the genetic model is known. The simulation was replicated 10000 times for the empirical power calculation. Table 2 shows that the power of the score statistic Z_x for a given genetic model is almost the same as that of T_{RLRT1} for the same genetic model and both are more powerful than T_{LRT1}.

When the genetic model is unknown, the empirical powers of MAX3, Z_{REC}, Z_{ADD}, Z_{DOM}, T_{LRT2} and T_{RLRT2}, were compared for four genetic models. The simulation was replicated 5000 times for the empirical power calculation. The critical values of MAX3 and T_{RLRT2} used in the power calculation are obtained by simulations as the upper α percentile of the empirical distributions of MAX3 and T_{RLRT2}. The results of simulated power under HWE are presented in Table 3. Table 4 gives the power comparison for a mixture of two populations with different allele frequencies $p^{**} = 0.20$ and $p^{**} = 0.05$.

In Tables 2 and 3, we also compared the power using the level of significance $\alpha = 0.01$. The conclusions, however, are similar to $\alpha = 0.05$. From Table 3 and Table 4, when the true genetic model is recessive (or dominant), Z_{DOM} (or Z_{REC}) has very low power. This is not surprising as the recessive and dominant models form two boundaries of the triangle T and have minimum null correlation among three score statistics Z_{DOM}, Z_{REC} and Z_{ADD} (Zheng *et al.* 2002). The TDT, Z_{ADD}, is more robust than Z_{DOM} and Z_{REC} across four genetic models. When the recessive model can be eliminated based on prior knowledge, Z_{ADD} is a robust test comparable to MAX3 and T_{RLRT2}. Overall, T_{RLRT2}, MAX3 and T_{LRT2} are robust across four genetic models. For the low allele frequency, MAX3 and T_{RLRT2} have similar power, but T_{RLRT2} is more powerful than MAX3 when the allele frequency is moderate ($p = 0.20$ or $p = 0.50$). It is also noted that T_{RLRT2} is always more powerful than T_{LRT2} in each situation studied. Similar results are obtained when the data consists of two subpopulations with different allele frequencies.

Discussion

In the case-parents design for testing candidate-gene association between a disease and a genetic marker, four genetic models (recessive, additive, multiplicative and dominant) are defined in terms of genotype relative risks. When the genetic model is correctly specified, score statistic is asymptotically optimal. For arbitrary genotype relative risks, we described a family of score statistics indexed by a parameter which is determined by a general genetic model. The score statistic that is optimal for the corresponding genetic model may have very low power when the genetic model is mis-specified, e.g., the optimal test for the recessive model has very low power when the true model is dominant and vice verse. In this situation, several robust statistics were studied, such as the maximum of three optimal statistics (MAX3) and the likelihood ratio test (LRT). Schaid (1999) compared the TDT (optimal for the additive model) to the LRT which is derived for a general genetic model and found that TDT is quite robust across four genetic models compared to the LRT. Zheng *et al.* (2002) compared score statistics and MAX3 across a broad family of genetic models and showed that MAX3 has efficiency

Table 3 Empirical power when the genetic model is known and HWE holds.

	True	Empirical power						
p	Model	(δ_0, δ_1)	MAX3	Z_{ADD}	Z_{REC}	Z_{DOM}	T_{LRT2}	T_{RLRT2}
			$\alpha = 0.05$ and $n = 150$					
0.05	Null	(1,1)	0.0458	0.0470	0.0176	0.0356	0.0362	0.0514
	REC	(0.4,0.4)	0.3642	0.1478	0.3276	0.0580	0.2544	0.3678
	DOM	(0.7,1)	0.6168	0.6946	0.0372	0.6518	0.4508	0.5604
	ADD	(0.6,0.8)	0.5084	0.5880	0.0788	0.4962	0.3396	0.4644
	MUL	(0.49,0.7)	0.6728	0.7242	0.1146	0.6883	0.5032	0.6556
0.20	Null	(1,1)	0.0468	0.0364	0.0296	0.0360	0.0362	0.0450
	REC	(0.4,0.4)	0.8508	0.4864	0.8994	0.1058	0.8014	0.8680
	DOM	(0.7,1)	0.4988	0.5348	0.0540	0.6158	0.4218	0.5648
	ADD	(0.6,0.8)	0.4458	0.5238	0.2032	0.5204	0.3754	0.5174
	MUL	(0.49,0.7)	0.6562	0.7628	0.3508	0.6658	0.5634	0.7158
0.50	Null	(1,1)	0.0550	0.0382	0.0478	0.0396	0.0492	0.0516
	REC	(0.4,0.4)	0.9502	0.8500	0.9660	0.0914	0.8842	0.9384
	DOM	(0.7,1)	0.6556	0.6412	0.1048	0.9014	0.7512	0.8650
	ADD	(0.6,0.8)	0.8052	0.8962	0.7380	0.7660	0.8194	0.9082
	MUL	(0.49,0.7)	0.6818	0.7994	0.6960	0.4842	0.6196	0.7686
			$\alpha = 0.01$ and $n = 200$					
0.05	REC	(0.4,0.4)	0.1996	0.0358	0.3586	0.0160	0.1390	0.2168
	DOM	(0.6,1)	0.7940	0.8582	0.0338	0.8910	0.7348	0.8358
	ADD	(0.6,0.8)	0.2640	0.3766	0.0676	0.3746	0.2162	0.3300
	MUL	(0.49,0.7)	0.4460	0.5742	0.1204	0.5600	0.3924	0.5212
0.20	REC	(0.4,0.4)	0.8338	0.4082	0.8282	0.0278	0.7408	0.8038
	DOM	(0.6,1)	0.7568	0.7396	0.0226	0.8258	0.6712	0.7648
	ADD	(0.6,0.8)	0.3354	0.3938	0.0924	0.3322	0.2392	0.3542
	MUL	(0.49,0.7)	0.5866	0.6628	0.2092	0.5394	0.4694	0.6140
0.50	REC	(0.4,0.4)	0.9996	0.9764	0.9994	0.0484	0.9946	0.9982
	DOM	(0.6,1)	0.2054	0.2840	0.0230	0.5520	0.3764	0.5066
	ADD	(0.6,0.8)	0.2408	0.3910	0.1950	0.2240	0.2388	0.3644
	MUL	(0.49,0.7)	0.5666	0.7184	0.5088	0.3390	0.5370	0.6608

robustness property compared to optimal statistics for each genetic model.

The restricted likelihood ratio test (RLRT) was applied to linkage analysis where IBD sharing probabilities are constrained to a smaller triangle (e.g., Holmans, 1993 and Knapp, 1998). We applied RLRT to the case-parents trio design for testing candidate-gene association, in which the genotype relative risks are constrained to a smaller triangle. The RLRT has a mixture distribution and the simulation results show that RLRT is more powerful than the LRT based on a general genetic model. Based on numerical results, across four genetic models, MAX3 and RLRT have similar power when the allele frequency is low and RLRT is more powerful than MAX3 for the moderate allele frequency. Hence, when the genetic model cannot be correctly specified, both RLRT and MAX3 can be used in case-parents design for testing candidate-gene association. For each statistic across four genetic models with minimum power, RLRT has the maximum power among these statistics studied. This indicates the efficiency robustness property of RLRT in case-parents design. MAX3 is relatively easier to calculate than RLRT, but both require simulation under the null hypothesis to obtain the critical values. When the recessive model can be excluded a priori, TDT is robust compared to RLRT, MAX3 and LRT and should be used. Moreover, as one referee pointed out that association tests based on genotype relative risks may give spurious result when stratified population is present, while TDT does not.

The research of Z. Li was supported in part by EY14478 of The National Eye Institute. We thank two referees for their valuable suggestions that improved our presentation. We also thank Drs. Nancy Geller and Eric Leifer for comments on the early version.

Table 4 Empirical power of test statistics for different genetic models and a mixture of two populations with different allele frequencies $p^* = 0.20$ and $p^{**} = 0.05$ ($\alpha = 0.05$ and $n = 150$).

True		Empirical power					
Model	(δ_0, δ_1)	MAX3	Z_{ADD}	Z_{REC}	Z_{DOM}	T_{LRT2}	T_{RLRT2}
Ratio of sample sizes of p^* to p^{**} is 2:1							
Null	(1, 1)	0.0488	0.0532	0.0294	0.0534	0.0414	0.0570
REC	(0.4, 0.4)	0.9200	0.5000	0.9200	0.1200	0.7200	0.8600
DOM	(0.7, 1)	0.8266	0.8628	0.0970	0.8848	0.7604	0.8602
ADD	(0.6, 0.8)	0.4698	0.5772	0.2062	0.4880	0.3510	0.5258
MUL	(0.49, 0.7)	0.6690	0.7642	0.3266	0.6636	0.5460	0.6988
Ratio of sample sizes of p^* to p^{**} is 1:1							
Null	(1, 1)	0.0478	0.0474	0.0300	0.0454	0.0438	0.0550
REC	(0.4, 0.4)	0.7064	0.3430	0.8156	0.0942	0.6080	0.7224
DOM	(0.7, 1)	0.8152	0.8710	0.0638	0.9008	0.7794	0.8554
ADD	(0.6, 0.8)	0.4302	0.5888	0.1524	0.5388	0.3644	0.5032
MUL	(0.49, 0.7)	0.6594	0.7614	0.3056	0.7280	0.5554	0.7034
Ratio of sample sizes of p^* to p^{**} is 1:2							
Null	(1, 1)	0.0478	0.0380	0.0294	0.0414	0.0422	0.0550
REC	(0.4, 0.4)	0.6410	0.2338	0.7670	0.0840	0.5086	0.6504
DOM	(0.7, 1)	0.8528	0.8706	0.1002	0.9234	0.7892	0.8812
ADD	(0.6, 0.8)	0.4556	0.5098	0.1802	0.5402	0.3528	0.5104
MUL	(0.49, 0.7)	0.6134	0.7320	0.2210	0.7128	0.5272	0.6976

Appendix A. Derivatives, Information Matrix and Mixture Proportion

Let the log-likelihood be $l = \log L_2(\delta_0, \delta_1)$. Then

$$\frac{\partial l}{\partial \delta_1} = (n_{21} + n_{41} + n_{51})/\delta_1 - n_2/(1+\delta_1) - 2n_4/(1 + 2\delta_1 + \delta_0) - n_5/(\delta_0 + \delta_1)$$

$$\frac{\partial l}{\partial \delta_0} = (n_{40} + n_{50})/\delta_0 - n_5/(\delta_0 + \delta_1) - n_4/(1 + 2\delta_1 + \delta_0)$$

$$\frac{\partial^2 l}{\partial \delta_1^2} = -(n_{21} + n_{41} + n_{51})/\delta_1^2 + n_2/(1+\delta_1)^2 + 4n_4/(1 + 2\delta_1 + \delta_0)^2 + n_5/(\delta_0 + \delta_1)^2$$

$$\frac{\partial^2 l}{\partial \delta_0^2} = -(n_{40} + n_{50})/\delta_0^2 + n_2/(\delta_0 + \delta_1)^2 + n_4/(1 + 2\delta_1 + \delta_0)^2$$

$$\frac{\partial^2 l}{\partial \delta_1 \partial \delta_0} = n_5/(\delta_0 + \delta_1)^2 + 2n_4/(1 + 2\delta_1 + \delta_0)^2.$$

The Fisher information matrix is $I = (I(i, j))_{2\times 2}$, where $I(i, j) = -\mathrm{E}(\partial^2 l/\partial\delta_i \partial\delta_j), i, j = 0, 1$. Evaluating I under the null hypothesis $\delta_0 = \delta_1 = 1$, we have $I(0, 0) = n_2/4 + 3n_4/16$, $I(0, 1) = I(1, 0) = -(n_2/4 + n_4/8)$, and $I(1, 1) = n_2/4 + n_3/4 + n_4/4$.

Write the Fisher information matrix under the null hypothesis as $I = Q\Lambda Q'$, a spectrum decomposition of I. Denote $Q = (q_{ij})_{2\times 2}$ and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$. Then $q_{21} = -q_{12}$ and $I(0, 0) = q_{11}^2\lambda_2 + q_{12}^2\lambda_1$, $I(1, 1) = q_{11}^2\lambda_1 + q_{12}^2\lambda_2$, $I(0, 1) = q_{11}q_{12}(\lambda_2 - \lambda_1)$ and. In the (δ_0, δ_1) plane, let the vertices of the triangle T be $O = (1, 1)$, $P_1 = (0, 1)$ and $P_2 = (0, 0)$. Then $OP_1 = P_1 - O = (-1, 0)$ and $OP_2 = P_2 - O = (-1, -1)$ are bases for the space T. From Self & Liang (1987),

$$\cos\delta = \frac{(P_1 - O)I(P_2 - O)'}{||\Lambda^{1/2}Q'(P_1 - O)'||\,||\Lambda^{1/2}Q'(P_2 - O)'||}.$$

It follows that $(P_1 - O)I(P_2 - O)' = n_4/16$, $||\Lambda^{1/2}Q'(P_1 - O)'||^2 = (4n_5 + 3n_4)/16$ and $||\Lambda^{1/2}Q'(P_2 - O)'||^2 = (4n_2 + 3n_4)/16$, which yield the left side of (6) as $q_i = n_i/n$. From Zheng *et al.* (2002), it equals to the null correlation of Z_0 and Z_1.

Appendix B. MLE Given a Genetic Model

The unrestricted MLE can be solved from the log-likelihood equation,

$$\partial \log L_2(\delta_0, g(\delta_0))/\partial\delta_0 = 0.$$

For the multiplicative model, $\hat{\delta}_1 = \hat{\delta}_0^2 = (n_{21} + n_{41} + 2n_{40} + n_{50})/(n_{22} + n_{41} + 2n_{42} + n_{51})$. For the recessive and dominant models, $\hat{\delta}_0$ are roots of

$$3(n_{22} + n_{42})\delta_0^2 + \{n_2 + 3n_4 - 4(n_{21} + n_{40} + n_{41})\}\delta_0 - (n_{21} + n_{40} + n_{41}) = 0,$$

$$(n_{41} + n_{42} + n_{51})\delta_0^2 + (n_4 + 3n_5 - 4n_{40} - 4n_{50})\delta_0 - 3(n_{40} + n_{50}) = 0,$$

respectively. For the additive model, $\hat{\delta}_1 > 1/2$ satisfies

$$6(n_{22} + n_{42})\delta_1^3 - (n_{21} + 5n_2 - n_{42} + 3n_{40} + n_{50} - 2n_{51})\delta_1^2 + (4n_{21} + n_2 - 4n_{42} - 2n_{40} - n_{50} + n_{51})\delta_1 - (n_{21} + n_{51} - n_{42} - n_{40}) = 0.$$

References

Holmans, P. (1993) Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet* **52**, 362–374.

Knapp, M. (1998) Evaluation of a restricted likelihood ratio test for mapping quantitative trait loci with extreme discordant sib pairs. *Ann Hum Genet* **62**, 75–87.

Kruse, R., Seuchter, S. A., Baur, M. P. & Knapp, M. (1997) The "possible triangle" test for extreme discordant sib pairs. *Genet Epidemiol* **14**, 833–838.

Self, S. G. & Liang, K. Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Statist Assoc* **82**, 605–610.

Schaid, D. J. (1999) Likelihoods and TDT for the case-parents design. *Genet Epidemiol* **16**, 250–260.

Schaid, D. J. & Sommer, S. S. (1993) Genotype relative risks: Methods for design and analysis of candidate-gene association studies. *Am J Hum Genet* **53**, 1114–1126.

Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506–516.

Zheng, G., Freidlin, B. & Gastwirth, J. L. (2002) Robust TDT-type candidate-gene association tests. *Ann Hum Genet* **66**, 145–155.

Received: 17 December 2002
Accepted: 23 July 2003